DOCUMENT RESUME

ED 418 138 TM 028 232

AUTHOR Burns, Matthew
TITLE Interpreting the Reliability and Validity of the Michigan Educational Assessment Program. Fact Finding on the Michigan Educational Assessment Program.
PUB DATE 1998-01-22
NOTE 15p.; Report of the Standing Committee for the Michigan Association of School Psychologists.
PUB TYPE Reports - Evaluative (142)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; *Achievement Tests; Criterion Referenced Tests; Elementary Secondary Education; Hispanic Americans; *Psychometrics; Standardized Tests; *State Programs; *Test Reliability; Test Use; Test Validity; *Testing Programs
IDENTIFIERS Michigan; *Michigan Educational Assessment Program

ABSTRACT

The psychometric properties of the testing tools of the Michigan Educational Assessment Program (MEAP), the state standardized testing program, are examined. Reliability studies have indicated that the scores from the MEAP, ranging from 0.654 to 0.949, are generally acceptable. The State Department of Education offered supporting evidence for the tests' criterion and construct validity and further concluded that no criterion evidence could be offered, since no other test matched the purpose of the MEAP. An independent evaluation by the Saginaw public schools (Michigan) suggested a generally low criterion validity for the story selection test and poor validity for Hispanic students. The remaining tests fell below an acceptable level for criterion validity. Another factor in considering the validity of the MEAP is the high stakes nature of the testing program, with its consequences for school districts. This may lower the validity of scores. Another factor is the exclusion of special education students. Overall, the MEAP has some advantages, but its tests have not demonstrated adequate reliability or validity to make decisions about individual students, asses writing skills adequately, assess Hispanic students, or make decisions about district or teacher effectiveness. Some alternatives to the MEAP's use are offered. (Contains 3 tables and 26 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 418 138

INTERPRETING THE RELIABILITY AND VALIDITY OF THE MICHIGAN EDUCATIONAL ASSESSMENT PROGRAM

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Matthew Burns

Fact Finding on the
Michigan Educational Assessment Program

Subcommittee on MEAP of the Standing Committee on Education

Presented by Matthew Burns
Regional Director
Michigan Association of School Psychologists

January 22, 1998

TMO28232

BEST COPY AVAILABLE

Interpreting the Reliability and Validity of the Michigan Educational Assessment Program

In 1989 Dr. Cannell, a psychiatrist and President of Friends for Education, surveyed the 50 states and was told that each scored "above the national average" on standardized testing. It does not take a psychiatrist to recognize the statistical impossibility of this occurrence. Since then, large-scaled standardized tests have come under great scrutiny. A similar movement has occurred in the field of school psychology as we examine our current standardized tools under the scope of psychometric properties, specifically reliability and validity. However, this critical information regarding state mandated standardized tests, while easily accessible to the public, is not common in the public knowledge. This paper will examine the psychometric properties of the testing tools from the Michigan Educational Assessment Program.

Reliability

The first aspect of a sound testing instrument is reliability, which has been defined as the consistency of test scores (Gronlund, 1993). It is a quantitative concept with an established level of adequacy. Reliability is often measured by conducting the test, and retesting the same sample with the same device a short time later, preferably two weeks (Salvia & Ysseldyke, 1991). The scores are then correlated with one another to establish the quantitative level. Ebel and Frisbie (1991) discussed the minimum level of reliability for a tool to be considered adequate. They listed .65 as the lowest acceptable level for a group test such as the MEAP, but also recommended using additional sources of information when using devices with reliability this low. Tests that are used to make educational decisions about individual students should obtain a

reliability level of at least .85. Tables one and two list the reliability scores of the MEAP tests as provided by the Michigan Department of Education.

These scores, ranging from .654 to .949, have general acceptance. The mathematics tests offer the most reliable scores since all of those coefficients fell above .90. Science tests all fell above .85, which also suggests reliability for those tests. However, the reading and writing tests demonstrated much more questionable reliability. Only Form B of the HSPT reading test fell above the minimum acceptance rate of .85. The remaining reading tests ranged from .674 to .842, and the writing tests scored at .654 and .674. These levels raise serious concerns about the data they generate. Inferences from these scores must be limited to group information. In other words, they offer no reliable data about individual students. Secondly, the tests that scored below the .70 range should be coupled with other data to make accurate assumptions about the groups they sampled.

Validity

The second issue addressed in an adequate testing device is validity, a much more complex and important concept. Generally speaking, validity refers to the extent that a test measures what its users claim it will measure (Salvia & Ysseldyke, 1991). There are three commonly accepted ways for a test to demonstrate its validity. Ebel and Frisbie (1991) described three common approaches: 1) Content validity, which demonstrates how well the content of a tool represents the domain of abilities the user is attempting to measure; 2) Criterion-related, which establishes relationships by correlating the test scores with a

Table 1

Reliability Data for MEAP Tests: Grades 4 and 7

	Grade	Component	Reliability
Reading			
Story	4	Constructing Meaning	.842
	4	Knowledge about Reading	.756
Informational	4	Constructing Meaning	.818
	4	Knowledge about Reading	.674
Story	7	Constructing Meaning	.790
	7	Knowledge about Reading	.781
Informational	7	Constructing Meaning	.792
	7	Knowledge about Reading	.771
Mathematics	4	NA	.934
	7	NA	.949
Science	5	NA	.875
	8	NA	.971

Table 2

Reliability Data for MEAP Tests: High School Proficiency Test

Test	Form	Reliability
Reading	A	.820
	B	.855
Mathematics	C	.911
Writing	B	.674
	C	.654
Science	C	.895

criterion measure of relevant abilities; and 3) Construct validity, or the scores meaning as a psychological construct.

The department of education uses the State Board of Education approved Michigan Essential Goals and Objectives for Mathematics Education, Reading Education, Science Education, and Writing Education to base its MEAP questions on. Specific test items are written by teams of Michigan teachers to be consistent with the objectives outlined by the State Board. A bias review committee then examines each test item for biases against any particular group. The Department of Education offers this approach as evidence of content validity, and this reviewer agrees with their claim. However, this sequence does not assure a match between what is tested and what is being taught in the classroom, a frequent criticism of standardized achievement tests (Salvia & Ysseldyke, 1991; and Travis, 1996). In addition, the Department of Education suggested that for a student to perform well on the MEAP tests, he or she must have "Mastered the entire domain, not just bits and pieces." (document provided by the Michigan Department of Education) However, standardized achievement tests are not considered a holistic approach, and have been called molecular and facts-based (Baker, 1991). In other words, standardized achievement tests do not accurately measure entire domains.

The Department of Education offered a four paragraph summary of the MEAP tests criterion and construct validity. The authors concluded that there was no other test that matched the purpose of the MEAP tests and therefore, no criterion evidence could be provided. To dismiss this concept based on such a premise, and to offer no evidence of construct validity, is psychometrically unforgivable. Saginaw Public Schools (1993) conducted their own evaluation

of MEAP tests based on criterion-related validity. Salvia and Ysseldyke (1991) outlined two forms of criterion validity (predictive and concurrent) and listed measures of each. In regards to achievement tests they recommended other achievement tests or teacher judgments of achievement as acceptable criteria. Grade point average was established as a representative of teacher judgement of achievement for the Saginaw Public Schools study. The overall correlations between MEAP scores and GPA were as follows: mathematics and GPA equalled a .551 coefficient, story selection and GPA measured a .365 correlation, and informational selection and GPA fell at a .465 level. According to MacEachron (1982), these levels offer only a questionable relationship. Therefore, if grade point average is an acceptable representation of teacher judgement of achievement, then the mathematics and informational selection tests demonstrate somewhat moderate validity. However, the story selection portion fell below the acceptable level and causes concerns about the test. Table 3 describes further criterion-related validity in regards to a racial breakdown. These results indicate a generally low criterion validity for the story selection test and poor validity for Hispanic students. Only the story selection test scored near a minimum level for acceptance when testing Hispanic students, and only marginally so. The remaining tests fell below an acceptable level. Mathematics exhibited an acceptable level for white and African-America students, and informational reading was borderline adequate for the same group. There are additional concerns about the MEAP tests validity not adequately addressed in quantitative terms. Messick (1989) described true validity as being derived from inferences about score meaning, and interpretation and implication for action. Large-scale standardized achievement tests are designed to make general Table 3

Criterion-Related Validity Coefficients for the HSPT and GPA

Test	White	African-America	Hispanic
Story Selection	.320	.408	.422
Informational Reading	.492	.463	.315
Mathematics	.535	.553	.377

interpretations about the results (Haladyna, 1992; Mehrens & Kaminski, 1989; and Nolet & Tindal, 1990), and are not meant to be interpreted for individual students. Two common uses of standardized achievement tests include evaluating teacher effectiveness (Hall & Kleine, 1990), and drawing district-to-district and state-by-state comparisons to establish success in educating students (Haladyna, 1992). However, these are not validated uses of the results (Berk, 1988; Guskey & Kifer, 1990; Haladyna, 1992; and Koretz, 1991), and to do so is an unethical practice that results in inaccurate assumptions.

Test score pollution

Cannell (1989) was among the first to point out that higher tests scores may not be the results of higher achievement. Instead he indicated it may be due to "cheating" as a result of demanding administrators, the public, and the media. These tests have become high-stakes devices since issues such as accreditation, high school graduation, teacher effectiveness, funding, and district effectiveness have been connected to the scores with negative consequences (Moore, 1994). Haladyna, Haas, and Nolen (1990) advanced the concept of "cheating" to develop the notion of test score pollution. They defined such pollution as any factor that distorts a standardized test score interpretation, an occurrence that can render the interpretation of the

scores invalid. There are three sources of documented pollution: test preparation activities, situational factors, and context (Haladyna, 1991). Included in this list are factors such as testwiseness training, curriculum matching, changes in the instructional program, presenting similar items to students (practice tests), test anxiety, stress, fatigue, motivation, test administration practices, language deficits, socioeconomic context, family influences, and excusing low-achieving students from taking the test (Haladyna, 1991).

It is human nature to compare oneself to his or her neighbors, but that is not the validated use of these tests. Comparing school districts, printing district results in the media, basing effectiveness judgements for teachers and administrators, and tying the results to funding lead to test score pollution. The Governor's recent warning to low scoring school district only intensified the high stakes nature of the MEAP tests, which in turn lowers the validity of the scores. The Department of Education, State Government, public, administrators, and teachers should cautiously interpret the results in a method validated for the given use.

Exclusion of special education students

A seemingly surprising factor related to test score pollution as defined by Haladyna (1991) is the exclusion of low-achieving students from taking the test. In Michigan, students participating in special education services and who receive 49% or less of their Reading/English instruction in a general education setting, have their results excluded from the school's summary results. In addition to invalidating the results, this is a questionable practice for several reasons.

Recent federal legislation has called upon states to hold all students to the same high expectations and to assure that they have the same educational opportunities (Bond, 1996).

Excluding special education students from MEAP testing is a violation of these mandates. In addition, it has been well documented that schools tend to concentrate resources on students who are included in their accountability standards (Theodore, 1996). Is the state of Michigan as interested in the effectiveness of programs for disabled students as they are those designed to teach the general education population? The current guidelines foster another dilemma of significant magnitude. As a school psychologist I participate in the team that determines special education eligibility for individual students. Over the past four years I have been approached on several occasions to test students for special education eligibility in order to exclude their scores from the school's summary report. Principals have also asked me to re-evaluate current special education students to justify increasing their time in a special education classroom over the 51% cutoff. My greatest personal concern has been with the disappointed responses from educators when notified that individual low-achieving students were not found eligible for special education. Out of respect for my current and former employers, I will not provide specifics, but let it be known that we have developed a system where it is preferable to diagnose a child with a disability than to have them participate in the state mandated assessment program.

The problem is as stated earlier, if accountability is the goal, then the current state mandated assessment program needs to be validated for that purpose, and the entire population must be included in the assessment. However, a predominate inadequacy of standardized tests is the assumption that all students can be assessed using the same instrument (Travis, 1996). To achieve accurate accountability, the current model would have to be radically revamped.

Conclusions

The Michigan Educational Assessment Program has some advantages. For example it is criterion, not norm referenced, and provides information that could improve instructional practices. However, standardized tests such as with the MEAP, are generally overused to fulfill state and local mandates, and the results are under used in serving instructional needs of teachers and students (Ebel & Frisbie, 1991). Specifically, the results and arguments listed above indicate that MEAP tests have not demonstrated adequate reliability or validity to make decisions about individual students, to adequately assess writing skills, to assess students of an Hispanic descent, or make decisions about teacher district effectiveness. The HSPT's stated purpose is to determine whether an individual student is eligible to earn an endorsement of the local diploma in specific content areas. This is a psychometrically unsound practice given the questionable reliability and validity of those tests. If the Michigan Department of Education is interested in making decisions about individual students, or if teacher school district accountability is a goal, then it is time to rethink the current assessment program.

The intent of this paper was to outline concerns about the MEAP tests from a psychometric standpoint. However, it is only fair to offer alternatives. One frequent answer to similar questions involves more authentic assessments, especially in difficult to assess subjects such as writing. Taking two samples of writing and assuming accuracy in determining writing proficiency is not possible. Instead, several samples from several days, in several modes are needed (Elbow & Belanoff, 1991). The Kentucky Education Reform Act (KERA) includes an assessment program involving student portfolios, and has been identified as a national example

of sweeping reform. Portfolio assessment allows for disabled students to participate in the program (Theodore, 1996), is an effective approach to improving student learning and measuring achievement (Seldin, 1991), address individual student differences (Travis, 1996), presents a broader sample of work that represents more typical behavior (Gronlund, 1993), and assures increased validity (Belanoff & Dickson, 1991). The KERA assessment program is not without its difficulties, but it has been shown to increase comfort and overall skill levels in writing tasks (Mincey, 1996), and is a viable option worth exploring. It was not the intent of this paper to outline the solution, but to demonstrate the need to find one. The Michigan Educational Assessment Program needs to be reformed to assure accuracy in its results, to enhance the validity of inferences made from those results, and to improve the education of all Michigan students.

The Subcommittee members should be commended for taking an interest in this crucial topic. There is significant work to be completed on the state-mandated assessment program, and I thank the committee for the time and commitment they have given today and elsewhere. Please do not hesitate to contact me if I may of any assistance in this venture.

References

Baker, E. L. (1991). Expectations and evidence for alternative assessment. Authentic assessment: The rhetoric and the reality. Symposium conducted at the annual meeting of the American Educational Research Association, Chicago, IL.

Belanoff, P., & Dickson, M. (1991). Portfolios: Process and product. Portsmouth NH: Boynton/Cook Publishers.

Berk, R. A. (1988). Fifty reasons why student achievement gain does not mean teacher effectiveness. Journal of Personnel Evaluation in Education, 1, 345-363.

Bond, L. A. (1996). Who's including students with disabilities in assessments? Counterpoint, 16, 7 & 12.

Cannell, J. J. (1989). How public educators cheat on standardized achievement tests. Albuquerque, NM: Friends for Education.

Ebel, R. L., & Frisbie, D. A. (1991). Essentials of educational measurement. Englewood Cliffs, NJ: Prentice Hall.

Elbow, P., & Belanoff, P. (1991). State University of New York at Stony Brook portfolio based evaluation program. Found in P. Belanoff & M. Dickson Portfolios: Process and product. Portsmouth NH: Boynton/Cook Publishers.

Gronlund, N. E. (1993). How to make achievement tests and assessments: 5th edition. Boston: Allyn and Bacon.

Guskey, T. R., & Kifer, E. W. (1990). Ranking school districts on the basis of statewide test results: Is it meaningful or misleading? Educational Measurement: Issues and Practice, 9, 11-16.

Haladyna, T. (1991). Generic questioning strategies for linking teaching and testing. Educational Technology, Research, and Development, 39, 73-81.

Haladyna, T. (1992). Test score pollution: Implications for limited English proficient students. Focus on Evaluation and Measurement. Proceedings of the National Research Symposium on Limited English Proficient Student Issues, Washington D.C.

Haladyna, T., Haas, N., & Nolen, S. B. (1990). Test score pollution. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Hall, J. L., & Kleine, P. F. (1990). Preparing students to take standardized tests: Have we gone too far? (ERIC Document Reproduction Service No. ED 334 249)

Koretz, D. M. (1991). State Comparisons using NAEP: Large costs, disappointing benefits. Educational Researcher, 20, 19-21.

MacEachron, A. E. (1982). Basic statistics in human services. Austin, TX: PRO-ED.

Messick, S. (1989). Validity. In R. L. Linn Educational Measurement: 3rd edition. Washington, DC: American Council on Education.

Mincey, K. (1996). The impact of KERA writing portfolios on first-year college writers. Paper presented at the Annual Meeting of the Conferences on College Composition and Communication, Milwaukee, WI.

Michigan Department of Education (1997). Reliability and Validity of the Michigan Educational Assessment Program (MEAP) Tests 1996-1997. Lansing, MI.

Mehrens, W. A., & Kaminski, J. (1989). Methods for improving test scores: Fruitful, fruitless, or fraudulent? Educational Measurement: Issues and Practice, 8, 14-22.

Nolet, V., & Tindal, G. (1990). Evidence of construct validity in published achievement tests. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Moore, W. P. (1994). The devaluation of standardized testing: One district's response to a mandated assessment. Applied Measurement in Education, 7, 343-367.

Saginaw Public Schools (1993). A correlational study into the relationship between grade point averages, Michigan Educational Assessment Program scores, and student absences for tenth grade Arthur Hill and Saginaw High School students, 1992-1993. Evaluation Report. Saginaw Public Schools, Saginaw, MI. (ERIC Document Reproduction Service No. ED 360 432)

Salvia, J., & Ysseldyke, J. E. (1991). Assessment. 5th edition. Boston: Houghton Mifflin Company.

Seldin, P. (1991). The teaching portfolio: A practical guide to improved performance and promotion tenure decisions. Bolton, MA: Anker.

Travis, J. E. (1996). Meaningful Assessment. The Clearing House, 69, 308-312.

Theodore, R. (1996). Alternate Portfolios allow Kentucky to assess all students. Counterpoint, 16, 1,6,& 12.

BEST COPY AVAILABLE